Opinion filtering in Social Networks

Lobna Azaza

Abstract

Customer satisfaction is the key secret of success for all industries. As people leave on the Web their opinions on products and services they have used, reputation of resources and services has become paramount. The democratization of the Web and social networks have facilitated the collection and the share of consumer opinion on their quality. These opinions enable suppliers to understand their expectations and needs. Specialized sites such as eBay and Amazon allow users to give their opinions on a variety of products and services.

However, openness and anonymity of the online opinion sharing communities makes the task of measuring the reputation very difficult. Users have different expertise levels and spammers joined the community with malicious behaviors. Therefore, it is important to filter opinions before any calculation of reputation. This work presents an approach for filtering collected opinions.

The proposed system reduces first the redundancy of opinions hidden behind different identifiers. It detects then the influences among the users in the case of shared opinions and promotes the most consistent profiles. The credibility measurement is based on a model of a heterogeneous social graph to capture the different relations between users, opinions and resources or services. Filtered opinions can be used then to calculate the reputation of services and Web resources.

In order to evaluate our framework performance, we start by modeling the filtering approach. Then, we valid the proposed approach through experiments. This is based on a random data generation and a variation of different criteria considered. Finally, suggestions on how the system performance can be improved are given.

Key words : Web resources, reputation, credibility, opinion filtering.

Dedications

Believing that there is no mountain higher as long as Allah is on our side,

I dedicate this thesis to my beloved parents,

my brothers and sisters for believing in me and encouraging me during all of my years of study.

I also dedicate it to all of my dear friends for their support.

Lobna

Acknowledgments

My sincere gratitude goes to **Pr. Rim FAIZ**, my master thesis supervisor for her encouragement and guidance that truly helped the progression of my research internship. I am also thankful for her availability, supervision, constructive criticism and valuable advices, allowing me achieving and improving this work.

The special thank also goes to **Pr. Djamal Benslimane**, the co-supervisor of this master thesis at University of Lyon 1, for accepting me among the team SOC and inspiring me throughout this research. I also thank him for his uninterrupted encouragement, time, and efforts.

With immense honor, I thank all of the LARODEC Tunis and LIRIS Lyon members for their help during my internship.

Finally, it is with gratitude that I thank all the jury members for agreeing to evaluate my work and all the professors for the knowledge and skills they gave me during all my years of study.

Contents

Introduction **1**

1 E-reputation and opinion mining in the Web **5**

1.1 Introduction 5

1.2 Social networks 6

1.3 E-Reputation 9

1.4 Opinion mining 13

1.4.1 Data mining 13

1.4.2 Web mining 16

1.4.3 Opinion mining 18

1.5 Conclusion 21

2 Problem description and state of the art **22**

2.1 Introduction 22

2.2 Problem description 23

2.2.1 Characteristics of Web data 23

2.2.2 Challenges about opinion data 24

2.3 State of the art 26

2.3.1 Multi-identifiers detection related works 26

2.3.2 Other works 27

2.3.3 Critics 29

2.4 Conclusion 30

3 The proposed approach: Opinion filtering 31

3.1 Introduction 31

3.2 Contributions 31

3.3 Multi-identifiers detection model 34

3.3.1 Motivations 34

3.3.2 Detection criteria 34

3.3.3 Description of the multi-identifiers detection model 36

3.3.3.1 Detection of multi-identifiers with email address/profile name criterion 37

3.3.3.2 Detection of multi-identifiers with opinion publication dates criterion 37

3.3.3.3 Detection of multi-identifiers with friends list criterion 39

3.3.3.4 Multi-criteria detection of multi-identifiers 39

3.4 Opinion data representation model 40

3.4.1 Aggregation of similar identifiers and elimination of redundancy 40

3.4.2 Representation of opinion data with the social graph 41

3.5 Credibility model 42

3.6 Conclusion 46

4 Experiments and results 47

4.1 Introduction 47

4.2 Algorithms 48

4.2.1 Detection of multi-identifiers with the criterion email-address . 48

4.2.2 Detection of multi-identifiers with the criterion profile name . 48

4.2.3 Detection of multi-identifiers with the criterion opinion publication dates 49

4.2.4 Detection of multi-identifiers with the criterion friends list . . 51

4.2.5 Multi-criteria detection of multi-identifiers 52
4.3 Experiments 53
4.3.1 Data description 53
4.3.2 Evaluation metrics 54
4.4 Results and evaluation 55
4.5 Conclusion 57

Conclusion **59**

Bibliography **61**

List of Figures

1.3.1 Importance of E-reputation 11
1.4.1 Web Mining Taxonomy 17

3.2.1 General architecture of the opinions filtering system 33
3.3.1 Example of a social network 36
3.3.2 Similarity matrix 38
3.3.3 Example of similar identifiers classes 39
3.4.1 Social graph of opinions data 42
3.5.1 Inter-evaluation 43
3.5.2 Direct Influence 44
3.5.3 Indirect Influence 44
3.5.4 Case of inconsistency 45

4.4.1 Results of criteria taken individually 57
4.4.2 Results of different combinations of criteria 57

List of Algorithms

4.1 Email-Address 48
4.2 Profile Name 49
4.3 Publication dates similarity 50
4.4 HAC Algorithm 51
4.5 Friends list similarity 52
4.6 Multi-criteria Similarity 53

List of Tables

1.1 Top 10 Most Popular Social Networking Sites 8

4.1 The approach results for 1000 identifiers 55

Introduction

It is not exaggerated to say that the World Wide Web had the most important impacts to the human life in the last years. It changed the way of doing business, providing education, and managing etc. Today, the Internet plays an increasingly important role and it has gradually infiltrated into every aspect of our lives because of its rich and varied resources. People are spending more time on the Internet in order to build some kind of large social entertainment community and try to make the relationship between members closer as they communicate with each other as frequently as possible.

Social networks brings people together in many inventive manners that were barely imaginable just a short time ago (David *et al.*, 2009). People are showing new forms of collaboration and communication, for example they are working, sharing, and socializing online. Besides, these new technologies play a vital role in the entrepreneurial actions and also help improve business models, and unlock numerous possibilities to study human interaction and collective behavior (Nicole *et al.*, 2007).

The World Wide Web is growing at an alarming speed in both size and types of services and contents. The democratization of Internet and social media have given rise to significant development very useful in different areas. Users have quickly taken an important role in the assessment of services and products and hence in their evolution thanks to technological progress, allowing them to exchange and publish information, rapidly and with a guarantee of anonymity.

Individual users are participating more actively and are generating vast amount of new data. These new Web contents include customer reviews and blogs that express opinions on products and services. As customer feedback on the Web influences other customer's decisions, these feedbacks have become an important source of information for businesses to take into account when developing marketing and product development plans.

The Web and social networks, whether they are open to all or developed in a professional context, form powerful tools widely used to promote expressing and sharing opinions. A large number of websites offer satisfaction surveys. Thus, the Web and social networks play a strategic role not only in systems reputation but also in the architecture of information systems. Social networks have certainly evolved. Their functionality is radically different from networks first appeared in the late 90s.

Expansion of the concept has favored the emergence and deployment of networks dedicated to the corporate such as LinkedIn[1] and Viadeo[2], and other thematic allowing users to share very specific topics. Particular networks can also be distinguished

[1] www.linkedin.com

[2] www.viadeo.com

like Twitter which is characterized by the brevity of messages exchanged, and the Facebook [3] network gathering an impressive number of users.

Therefore, the trust of a user or a community of users, and e-reputation of products have become an important issue. This trust measurement is mainly based on the users opinions. The quality of collected opinions is undoubtedly a major challenge to better appreciate the confidence that can be given to a product or service.

Unfortunately, opinions data in applications that use social networks can be polluted by users in different ways. Individuals have sometimes malicious behaviors in order to promote or degrade the reputation of a product or service. Such tools in an open context raise real questions about the quality of collected opinions. Reputation assessment based on opinions poses two main challenges. The first is the use of different identifiers by the same user (Frederik *et al.*, 2013 ; Tieyun et Bing, 2013 ; Hung-Ching *et al.*, 2004 ; Arjun *et al.*, 2013). The second is related to credibility of users who express opinions (Yi-Cheng *et al.*, 2012). The more users are credible, the better their opinions affect reputation measurement (Metzger, 2007).

Therefore, it is important to filter opinions before any reputation measurement. The objective is to establish a collection of opinions that are filtered so that all forms of redundancy are eliminated and user credibility is given. Filtered opinions can be used then in works that aim to measure online resources reputation (Mostafa, 2013 ; Dingding *et al.*, 2013).

[3] www.facebook.com

In this work, we propose different contributions related to the objective of opinion filtering in an open environment threatened by malignancy. We offer an architecture for filtering opinions to improve their quality before calculating the product's reputation, we propose first a model for detecting virtual users that correspond to the same physical user. Then we suggest a model for calculating credibility of users that takes into account different criterion deduced from a heterogeneous social graph.

Document organization

This document is organized as follows:

- Chapter 1 reviews the background of e-reputation and opinion mining in the Web as the framework for this work.
- Chapter 2 describes the problem of opinion mining in the Web and presents the essential work of the state of the art in order to solve problems related to this issue.
- Chapter 3 describes the mechanism of the proposed approach: opinion filtering.
- Chapter 4 validates the proposed approach through experiments and presents the results of experimental evaluations.

We finish by concluding the research work proposed, giving guidelines for future work, and opening questions recently emerging in these areas.

Chapter 1

E-reputation and opinion mining in the Web

1.1 Introduction

The past decade has witnessed a rapid development and change of the Web and the Internet, social networks brings people together and users have quickly taken an important role in the assessment of services and products. Hence, e-reputation and opinion mining have become an important issue.

Previously, we mentioned some details of our problem and contributions. Before going to deep details, we introduce in this chapter a general overview of the framework.

This chapter is organized as follows: section 1.2 presents the Social Networks field. Section 1.3 defines the e-reputation. And finally section 1.4, describes opinion mining.

1.2 Social networks

The Web and the Internet are growing at an alarming rate, today they play an increasingly important role because of its rich and varied resources. Web applications and social networking sites have been cropping up, gathering people together and empowering their relationships with new forms of cooperation and communication (Danah and Nicole, 2007).

A social network is a theoretical concept in the social sciences, particularly sociology and anthropology, referring to a social structure made up of individuals or organizations[1]. Social networking is most known online, despite the fact that it is possible in person, especially in workplaces.

This is due to the fact that Internet is very different from ordinary social networks like schools or universities. The Internet is used by millions of persons who aim to meet other users in order to collect and share information and experiences about various topics and interests. Information might be about cooking, study, traveling, business etc.

Websites are popularly used when we are interested in online social networking, more known as social sites. Social networking websites are like an online unity that brings Internet users closer together. You simply need an access to a social networking website to start socializing. The content of the shared information depends on the website itself, many of the online community members show common interests

[1] http://en.wikipedia.org/wiki/Social_network

in hobbies, religion, politics and alternative lifestyles.

As mentioned, social networking often implies gathering specific users or associations together. While there are social networking websites which focus is on specific concerns, there are others that do not have any particular topic. They're called traditional social websites as they don't fix a main focus and memberships are open to all.

In traditional social networking websites, anyone can become a member, regardless their hobbies, beliefs, or views. Yet, once you are a member of this kind of online community, you are able to start creating your own network of friends and exclude others that do not share common interests or goals.

Whether traditional or not, Social networks provide a strong glint of the society structure of the 21st century as there are hundreds of social networking sites upholding a large variety of interests and practices.

Table 1.1 lists the top 10 Most Popular Social Networking Sites as derived from eBizMBA Rank[2] which is a continually updated average of each website's Alexa Global Traffic Rank[3], and U.S. Traffic Rank from both Compete[4] and Quantcast[5].

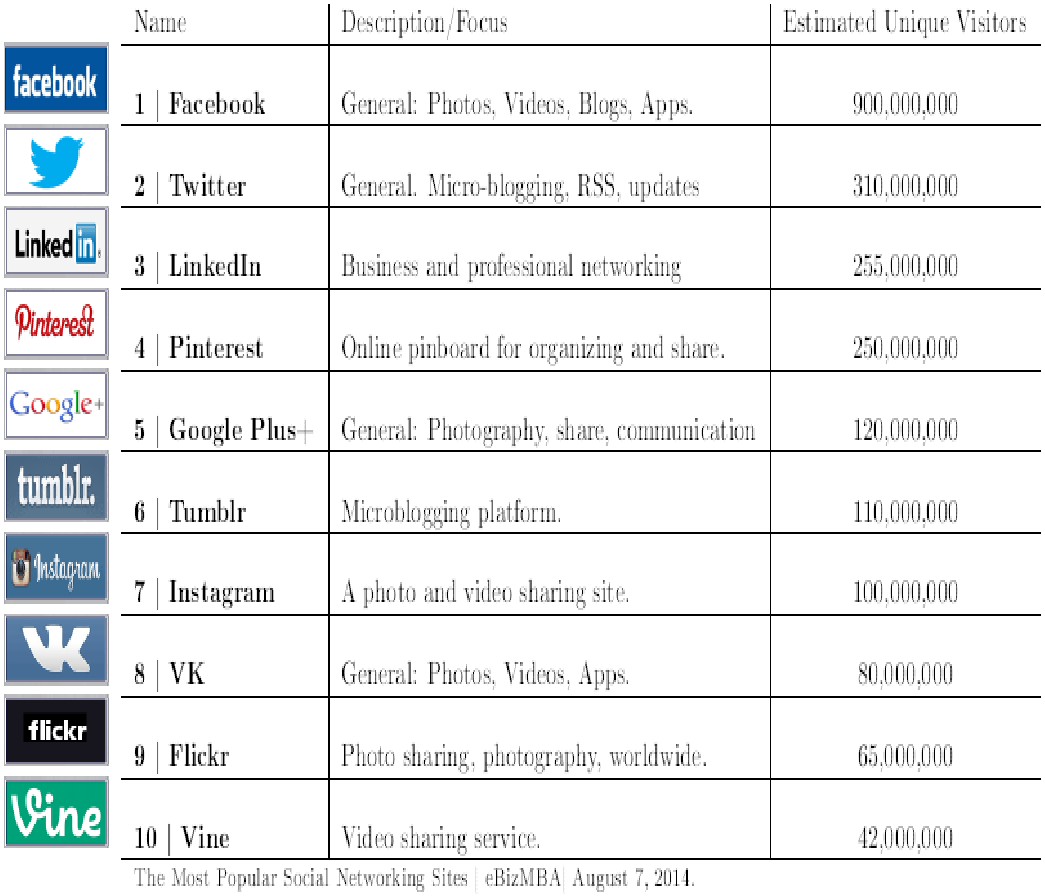
[2] http://www.ebizmba.com/articles/social-networking-websites

[3] http://www.alexa.com/topsites

[4] https://www.compete.com/

[5] https://www.quantcast.com/top-sites

Table 1.1: Top 10 Most Popular Social Networking Sites



Name	Description/Focus	Estimated Unique Visitors
1 | Facebook	General: Photos, Videos, Blogs, Apps.	900,000,000
2 | Twitter	General. Micro-blogging, RSS, updates	310,000,000
3 | LinkedIn	Business and professional networking	255,000,000
4 | Pinterest	Online pinboard for organizing and share.	250,000,000
5 | Google Plus+	General: Photography, share, communication	120,000,000
6 | Tumblr	Microblogging platform.	110,000,000
7 | Instagram	A photo and video sharing site.	100,000,000
8 | VK	General: Photos, Videos, Apps.	80,000,000
9 | Flickr	Photo sharing, photography, worldwide.	65,000,000
10 | Vine	Video sharing service.	42,000,000

The Most Popular Social Networking Sites | eBizMBA| August 7, 2014.

Social networks have become widespread in the age of the Web thanks to interfaces that allow people to follow their friends lives, knowledge and families, the number of social networks users has grown exponentially since this century's beginning.

For instance, Facebook and Twitter[6] have attracted millions of users, many of

[6] www.twitter.com/

whom have integrated these sites into their daily practices. Such networks provide a highly appropriate frame to share information between individuals and their friends in an instantaneous manner.

Indeed, the dramatic raise of social networks and user generated content is influencing all stages of the content value process including production, processing, distribution and use. It also created and introduced to the multimedia area a different critical look of science and technology which is social interaction and networking.

This fresh speedily growing research area is very important and this is justified by the many associated evolving technologies and applications including online content sharing services and communities, multimedia communication over the Internet, social multimedia search, interactive services and entertainment, health care and security applications. This leaded to the emergence of the social multimedia computing, in which well established computing and multimedia networking technologies are brought together with emerging social media research.

1.3 E-Reputation

The rise of social networks (or social media) over the past few years, figures among the most important Web phenomena. Nowadays companies believe that social networks are the most important social media channel and some of them even spend more time concentrating on their Facebook page than on their site. Today's marketers cannot ignore the worth of this phenomena and must make strategic decisions including their presence on social networks.

Social networks affected different functions of the company especially reputation management in particular. This is not surprising as we are in the era of self-publishing, nowadays publishing is given to everyone all over the world at zero cost. It is no longer necessary to follow a costing process such as publishers and printers to send messages to people.

Distribution also became free thanks to the social network phenomenon. In addition to self-publishing, individuals are able to connect with one another easily,instantaneously and again at zero cost. Which means that publishing and distribution are now free and open to all.

With such conditions, we easily think that this phenomenon offers a dream opportunity for all to communicate and share ideas and opinions on products and services. But, as was never the case before, a brand's reputation depends on several factors, a majority of which the brand itself is no longer able to control.

It has become a key priority to catch different opinions that Internet users write in social networks. Those opinions are useful to measure a brand's reputation on the Web, as reputation is what people say about the product.

It is possible that no one talks about a particular brand which would be quite dramatic, but there are more chances that brands have an important number of messages in articles, blogs, tweets, comments etc. Also technologies are well developed to detect all these data rich of information.

Online reputation (E-reputation) today matters more than ever. A lot can happen to a company's online reputation when it's not watching. Actually, this kind of conversations is going between past consumers and maybe future ones online on review sites or social networks. Companies should not take the risk of not knowing what is happening and said about their brands.

Figure 1.3.1: Importance of E-reputation

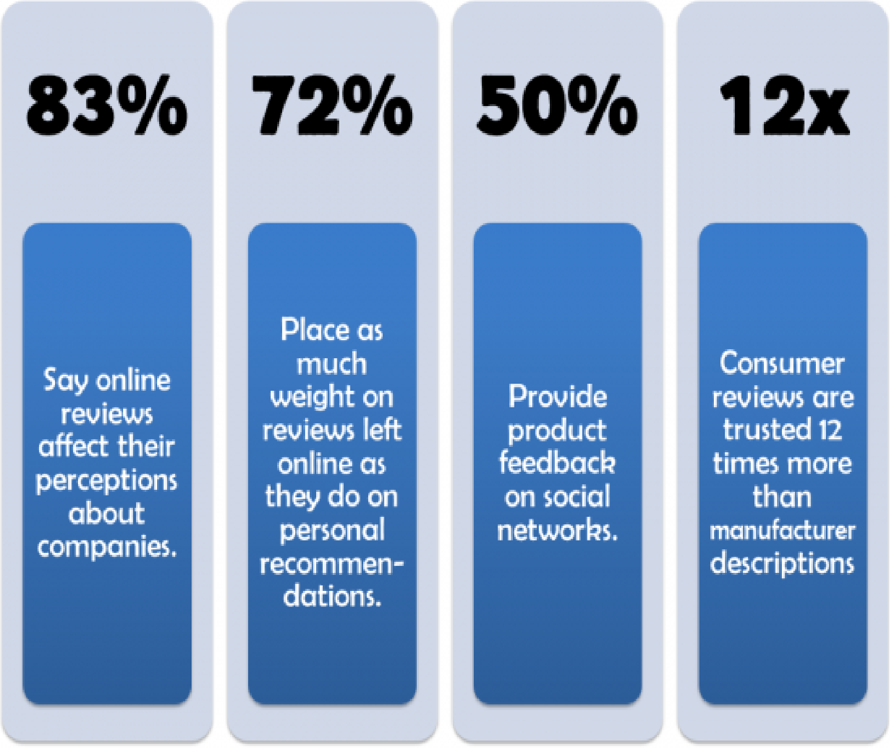


Figure 1.3.1 shows recent statistics which provide more evidence that every com-

pany should take a closer look at its online reputation. WebProNews[7] announced that 83% of consumers say online reviews affect their perceptions about companies: Opinions are simple to post, easily found online, and affecting the behavior of more than 8 out of every 10 reviewers.

In the other hand, according to Search Engine Land[8], 72% of consumers said that they “trust online reviews as much as personal recommendations”. Trust is an important issue especially online, yet, almost 3/4 of customers have faith in online feedbacks as much as they do on friends and family’s recommendations.

Also, Monetate[9] shows that 50% or more use social networks to provide product feedback, whether it is positive or negative. Users are aware that this way their voices are well heard and taking advantage of this opportunity.

Finally statistics show, according to a survey of US internet users, that product feedbacks are trusted 12 times more than descriptions from manufacturers (Econsultancy[10]). Information provided by manufacturers on their websites is important, but today a past customer is more able to convince future prospects.

As the Web evolved to give rise to the social networks, technology has followed advancement and marketers have access to several and powerful tools to catch huge volume of data containing e-reputation information. It is no longer a question of simply parsing a few press articles, but rather detecting what is said about a brand

[7] http://www.webpronews.com/
[8] http://searchengineland.com/
[9] http://www.monetate.com/
[10] https://econsultancy.com/

all over the Web, how it is said and how it can influence its online business.

Despite the overflow of easily accessible messages expressed by Internet users, for some people the debate on whether there is a link between this activity and the success or failure of an online business remains open. No one believes that a poor online reputation does not have an ultimate impact on a company.

However, any negative comments will not have necessarily dramatic consequences for a brand. Monitoring skills should not interpret any negative feedback as a source of panic. But, it is important to recognize that something published on Twitter, Facebook or on blogs today has great effects on consumers even if it does not necessarily have the same impact as if it were published in a newspaper or in books.

1.4 Opinion mining

It is important to figure out that e-reputation is associated with the sentiment analysis or opinion mining, which is data mining where the content of data is opinions. In other words, the techniques of measurement of e-reputation or determining if Internet users speak highly or poorly of a brand.

1.4.1 Data mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), is an interdisciplinary sub-field of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. [11]

The data mining task is to automatically analyze large volume of data in order to extract patterns previously unknown like data groups (clustering), unusual behaviors (anomaly detection) and dependencies (association rule mining).

Patterns discovery involves using database techniques. As a summary of the input data, these patterns can be used then in analysis process such as machine learning and prediction. Data mining can make a decision support system more solid by giving more accurate results. For example it can detect multiple clusters in the data that can be used in further prediction tasks.

Data mining and KDD should not be confused together. Data mining is a basic step of KDD, which means that data selection, transformation and result evaluation are not part of the data mining process but belong to KDD procedure.

The KDD process is commonly simplified as pre-processing, data mining, and results validation.

Pre-processing is accomplished before any use of data mining algorithms. Input data should be chosen first in a way that it is concise and huge enough enabling data mining process to discover patterns. This can be done by data warehousing tools. Noise and anomalies are also removed from data to have the target clean and ready

[11] http://en.wikipedia.org/wiki/Data_mining

to be used in data mining process.

Data mining involves six common categories of tasks (Fayad *et al.*, 2008):

- Anomaly detection – The recognition of unusual data behavior, that might be interesting or data errors that require further investigation.
- Association rule learning – Looking for dependency relationships between variables.
- Clustering – The task of discovering similar groups and clusters in the data without previously known structure.
- Classification – The task of generalizing previously known structure to apply to new data.
- Regression – Finding a function which models the data with the least error.
- Summarization – Representing the data set in a more tight way.

The final step of knowledge discovery from data which is results validation consists on verifying whether the patterns produced are valid or not. It is common for the data mining algorithms to find patterns that are not similar to desired standards. This is known as overfitting. To avoid this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set, and the resulting output is compared to the desired output.

If the learned patterns are not identical to the desired standards, it is necessary to re-evaluate and change the first and the second steps until having the desired standards. If so, the final step is to interpret the learned patterns and turn them into knowledge.

1.4.2 Web mining

Web mining is a fresh, hot and very promising research discipline. It combines two of the most important research fields which are World Wide Web and Data mining. By mixing both of those topics, Web Mining became one of the most popular disciplines in Web and got into pool of interest for many researchers.

The term of Web Mining has been proposed first by Oren Etzioni (Oren, 1996). It is the fact of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services.

Despite the fact that Web mining lay the roots profoundly in data mining, it is not parallel to data mining. Web data raises up more complexity of Web mining.

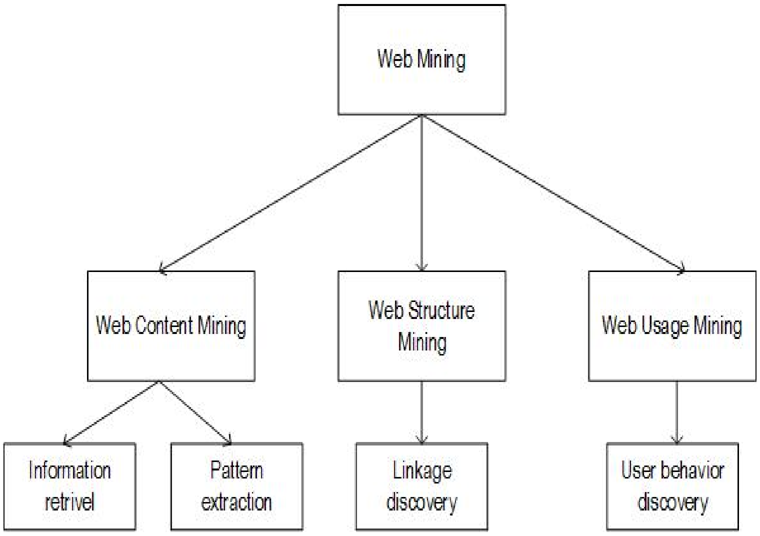
The Web mining research is actually linked to many other research fields, it is a converging area from communities such as Database, Information Retrieval, Artificial Intelligence, and also psychology and statistics as well (Raumond and Hendrik, 2000).

There exists quite some confusion about this research field. Yet categorization of Web mining stills the most admitted method. Thus, As shown in figure 1.4.1, Web mining can be classified into 3 categories based on which part of the Web is to be mined: Web content mining, Web structure mining, and Web usage mining.

Web content mining emphasizes to useful information discovery/retrieval from the Web, whereas Web structure mining is the process of extracting knowledge from the interconnections of hypertext document in the world wide web. For example using linkage information to improve search engines.

The distinction between Web content mining and Web structure mining is quite hard and fuzzy sometimes. It is not the case with the Web usage mining which is relatively independent, but not separated. It focuses on user's usage pattern discovery and behavior prediction.

Figure 1.4.1: Web Mining Taxonomy



The goal of the Web content mining concerns a large variation of applications which purpose is to discover and extract hidden information in data stored on the Web. Whereas, the target of the Web structure mining is to provide a mechanism to make the data access more efficiently and adequately. And finally Web usage mining aims to discover the information that can be derived from the users activities, which are stored in log files for example for predictive Web caching.

While web structure and content mining use primary data on the web, web usage mining works on the secondary data such as web server access logs, proxy server logs, referrer logs, browser logs, error logs, user profiles, registration data, user sessions or transactions, cookies, user queries, and bookmark data.

1.4.3 Opinion mining

Opinion mining belongs to Web content mining as it is the process of tracking the judgment of the public about a certain product, brand or a service.

Opinion mining is also called sentiment analysis. It includes constructing a system able to extract and classify opinions about a brand or service. This means the measurement of a product's e-reputation by determining if Internet users speak positively or negatively about a brand. Messages that have been captured (negative, positive or neutral) have to be sorted into the different categories, and here lies a major defiance. The field often uses a type of artificial intelligence which is machine learning in order to elicit text for sentiment.

Opinion mining is beneficial in various ways. It can help marketers assess the influence of an ad campaign or new product launch on Internet users. For example, a feedback on a website might be positive on the whole about a mobile phone, but be specifically negative about its camera quality. Identifying this sort of information systemically enables having a bright image of public opinion that can't be achieved with surveys or focus groups as evaluations are spontaneously given by the customer.

However, such domain faces several challenges. The first one is related to the different meanings that a word can have. Actually it depends on the situation and the product itself. A word can be considered positive in a case and negative in another. For example the word "big", a consumer can say that a laptop has a big memory which can be considered as a positive feedback. But if he says that the weight of the laptop is big, this is a negative opinion. Hence, an opinion mining system should be trained on this and perform rightly whatever the type of product is.

Another challenge lies in the different ways with which people express their ideas. In expressions of opinion and judgment, a little difference between two sentences can change radically its meaning. "I recommend this product" is totally the opposite of "I don't recommend this product".

The last challenge is about contradiction that customers can have in their assessments. Which means that a review can be together positive and negative. It is easy for a human to figure out the meaning of a sentence when the writer combines a negative and a positive expressions at the same time. However, it is more difficult for a machine to understand it. For example, “This product is very good although medias said that its functionality is limited”. Sometimes even a human can’t understand the meaning of an expression like “The product is as good as another product”. This depends on the customer’s opinion about the other product which makes the task harder for the machine.

Many solutions are available to automatically extract opinions and measure the e-reputation. They differ highly in their ability to find out whether an article or a tweet expresses a negative, positive or even neutral opinion. It depends on how deep systems consider the challenges listed above.

Once the feedbacks have been classified into their categories, it is possible to measure the e-reputation. Here are two simple calculations which can be used:

1. Percentage of positive evaluations: it implicates taking the total number of positive evaluations and dividing this figure by the total number of evaluations.

2. Ratio of negative evaluations: this calculation measures negativity. It concerns finding the number of negative feedbacks existing for each positive message. For a ratio of 1:2 means that for every positive evaluation there are two negative ones, this would surely create a state of panic.

1.5 Conclusion

In this chapter, we presented briefly the background related to our research framework. This included social networks, e-reputation and opinion mining. We mentioned how important became this area and how the research topic is very hot. This domain faces several challenges which we discuss in the next chapter.

Chapter 2

Problem description and state of the art

2.1 Introduction

Social media explosion has reproduced unexpected opportunities for users to express and publish their opinions. Unfortunately, this phenomenon has witnessed serious problems when it comes to using and making sense of these feedbacks. This happens while the gain of a real-time understanding of customers needs is growing and became a very urgent task.

Policy-makers and people are trying to build effective systems to make use of this huge data about meaningful interactions between thousands of social networks users. We are therefore at a crucial position where the challenge of information overload can be transformed from a problem to an opportunity for making sense of a large amount of opinion data.

In this chapter, we are going to detail the main problems about Web data in general, then opinion data specifically. And we finish by giving some previous works who attempted to deal with these bottlenecks.

2.2 Problem description

2.2.1 Characteristics of Web data

With the intensive alerting and explosive development of information accessible over the Internet, World Wide Web has become a strong platform to stock, share and discover information as well as extract useful knowledge.

Due to the large, various, dynamic and unstructured characteristic of Web data, research in this field has confronted a lot of challenges, such as heterogeneous structure, distributed residence and scalability issues etc. In fact, Web users encounter a problem of overloaded information when using the Web and find themselves lost in an ocean of information. Typically, the following characteristics are often related to Web data and cause a difficulty in Web researches and applications.

First of all, data on the Web is huge in amount. At the moment, it is hard to estimate the exact data volume available on the Internet due to its exponential growth every day. The enormous volume of data on the Internet makes it difficult to well handle an explore Web data.

Second, Web data is distributed and heterogeneous. Due to the fact that the Web is a meeting point of various nodes over Internet, Web data is generally distributed through a wide range of computers or servers, which are situated at different places around the world. Meanwhile, Web data is often displaying the fundamental characteristic of multimedia, which is, besides textual information, that is mostly used to express content in terms of text message; many other types of Web data, such as images, audio files and video are often included in a Web page.

Also, data on the Web is unstructured. Till the moment, there are no uniform data structures or schemes that should be strictly followed by Web pages. Web designers have the ability to randomly organize connected information on the Web together however they want, without following specific ways, as long as the information arrangement meets the basic layout requirements of Web documents, such as HTML format.

Finally, Web data is dynamic. The implicit and explicit structure of Web data is often updated. Notably, Web based database system application has caused the fact that a variety of presentations of Web documents are generated as contents in database update. Indeed, domain or file names changes or disappear which will produce problems of handled links and relocation.

2.2.2 Challenges about opinion data

Opinion data are similar to Web data in characteristics perspective. Besides the features that we described in the previous section, there are other ways in which opinion data can be infected.

Opinion data especially in applications that use social networks can be polluted by users in different ways. Individuals have sometimes malicious behaviors in order to promote or degrade the reputation of a product or service. For example, they can evaluate the same product several times from their accounts or hiding behind multiple identifiers (Frederik *et al.*, 2013 ; Tieyun et Bing, 2013 ; Hung-Ching *et al.*, 2004 ; Arjun *et al.*, 2013). This leads to obtain redundant opinions.

Evaluators can also give opinions about products or services that they haven't experienced by themselves but rather reflecting their social friends opinions (Md Yusuf *et al.*, 2012).

In addition, products evaluators have different expertise and the fact of taking into account all evaluations without considering their credibility can distort the calculation of the products reputation (Yi-Cheng *et al.*, 2012 ; Guan *et al.*, 2011).

Although useful, such tools in an open context raise real questions about the quality of the opinions collected. The trust measurement based on opinions gathering poses two main challenges.

- The first challenge is about using different identifiers by the same physical user. Such concealment of a user aims to disrupt the collection of opinions to impact the measurement of the products e-reputation. It is therefore important to identify the virtual users that might correspond to a single physical user to filter the collected reviews, eliminate the redundant opinions and only keep the most significant ones.

- The second challenge is related to the users credibility who express opinions. It is related to the degree of credibility of the users, and so their opinions. The more the users are credible, the better their opinions will affect the reputation measurement (Metzger and Miriam, 2007). It is then important to be able to estimate the credibility of users and update it while collecting opinions.

2.3 State of the art

2.3.1 Multi-identifiers detection related works

Regardless of the type of social network used, the discovery of the same physical user behind different user identifiers is a current and important issue.

When it comes to multi-identifiers detection, the main work related to ours is the one described in the paper of Frederik et al. (2013) in which authors propose a number of techniques for alias matching: string based, stylometric-based, time profile-based, and social network-based matching. If the idea is attractive, unfortunately the results given by this approach are not satisfying. They do not offer enough combinations of techniques proposed for users comparison.

Our approach aims to combine various criteria in order to build a more plausible multi-identifiers detection system. Also we don't consider features related to linguistic tools. And features proposed are used to calculate a similarity score between users while they only propose the techniques and keep the choice of their combination open.

A similar problem has been studied by Tiyeum and Bing (2013), they propose a new method based on linguistic analysis to identify userids that may be from the same author. Our challenge is to recognize identifiers of the same user without using any linguistic tools.

Hung-Ching et al., (2004) present an algorithm to identify multi-identifiers users, which is based on a model of communication exchange on a public forum. They observed that the posts of an identifier operated by a multi-identifiers actor do not appear as frequently as do the posts of single-identifier users. All posts of multi-identifiers users are correlated but do not occur too close together. The algorithm detects the identifiers whose posts display such statistical anomalies and identify them as coming from multi-identifiers users.

One of the other methods proposed in literature lies in the analysis of the content published by virtual users (Nitin and Bing, 2008). This analysis aims to calculate the similarity of vocabularies. Such method can be efficient in a particular field but shows its limits in the case of different areas.

The approach proposed by Arjun *et al.* (2013) measures the similarity between the virtual users regardless of subject matter. The proposed approach is based on a set of characteristics.

2.3.2 Other works

Various researches were interested in the diversification of sources of evaluations in the process of calculating e-reputation. Thus, the system TIDY (Anthony *et al.*, 2013) defines the concept of diversification by aggregating similar sources into a single virtual source. Such aggregation is based on similarity metrics and learning techniques.

Meanwhile, the system Truth Finder (Xin Luna *et al.*, 2009) also uses similarity metrics between sources but rather to choose those that are less similar, and therefore the most representative of all evaluators.

Other approaches define the diversification of sources with focusing on sources that interact less with each other, and thus minimize the influences that they may have on each other (Md Yusuf *et al.*, 2012).

Some approaches have considered the evaluators credibility through the exploitation of various types of information for estimating the expertise relative to the subject of study. In fact, Tracy and Robert (2001) simply associate the evaluators credibility with their expertise. This can be derived from exploiting their publications.

Other systems are based on the idea that the reputation and credibility of the evaluator must be based on the evaluation of evaluations. The more user ratings tend towards consensus and majority opinions, the more evaluators gain in credibility. Algorithms are proposed for a continue update of the evaluations (Yi-Cheng *et al.*, 2012).

The field of Web services has also benefited from the researches on reputation. However, Zaki and Athman (2009) propose the selection of the best services by analyzing the feedback from users and measuring the distance between the feedback and the majority opinion. The majority opinion is represented by the centroid of the most populous group generated by the K-mean algorithm on feedbacks listed as being quite similar.

In the paper of Zohra *et al.* (2014), the notion of severity is proposed to less penalize the credible users but unfortunately quite far from the majority opinion. The algorithm of Fuzzy C-mean Clustering is implemented.

2.3.3 Critics

The research field concerned in this thesis is a very hot topic nowadays. Literature has witnessed many works that attempt to facilitate exploration of opinion information in the Web.

Unfortunately, these works represent some limits which are different from an approach to another.

Some proposed systems resort to linguistic features (Frederik *et al.*, 2013; Tieyun and Bing, 2013; Nitin and Bing, 2008), which is very complicated because sometimes, even a human is unable to understand the meaning of a feed-back. This make the automated language processing based systems far away from providing the good results.

Other works don't use linguistic tools and rather exploit other features. But most of them use only one criterion which is not enough. Communication exchange (Hung-

Ching *et al.*, 2004) or inter-evaluation (Yi-Cheng *et al.*, 2012) can not be sufficient to evaluate a user's credibility. Others evaluate credibility as the expertise level of the user (Tracy and Robert, 2001).

Another interesting area has been considered which is diversity (Anthony *et al.*, 2013 ; Xin Luna *et al.*, 2009 ; Md Yusuf *et al.*, 2012), it can be interesting to eliminate redundancy based on some characteristics. But remains insufficient.

All these characteristic or features are interesting, but none of the works has proposed an approach in which they combine different criteria for both multi-identifiers detection and credibility measurement. Also there is not a system that takes into consideration these two issues in the same work.

2.4 Conclusion

In this chapter, we described the different problems faced by systems who use opinion data in the Web. We also gave the different approaches proposed in the literature in order to solve these problems.

In the next chapter, we detail our approach titled opinion filtering which includes solutions for both multi-identifiers detection and users credibility measurement.

Chapter 3

The proposed approach: Opinion filtering

3.1 Introduction

Data in applications that use social networks have characteristics that make their exploration difficult. There are some previous works that attempted to solve this issue but they present some limits.

In this chapter, we present our approach for opinion filtering. The system reduces first the redundancy of opinions hidden behind different identifiers. It detects then the influences among users in the case of shared opinions and promotes the most consistent and reliable profiles.

3.2 Contributions

In this work, we propose different contributions related to the purpose of opinion filtering in an open environment threatened by malignancy.

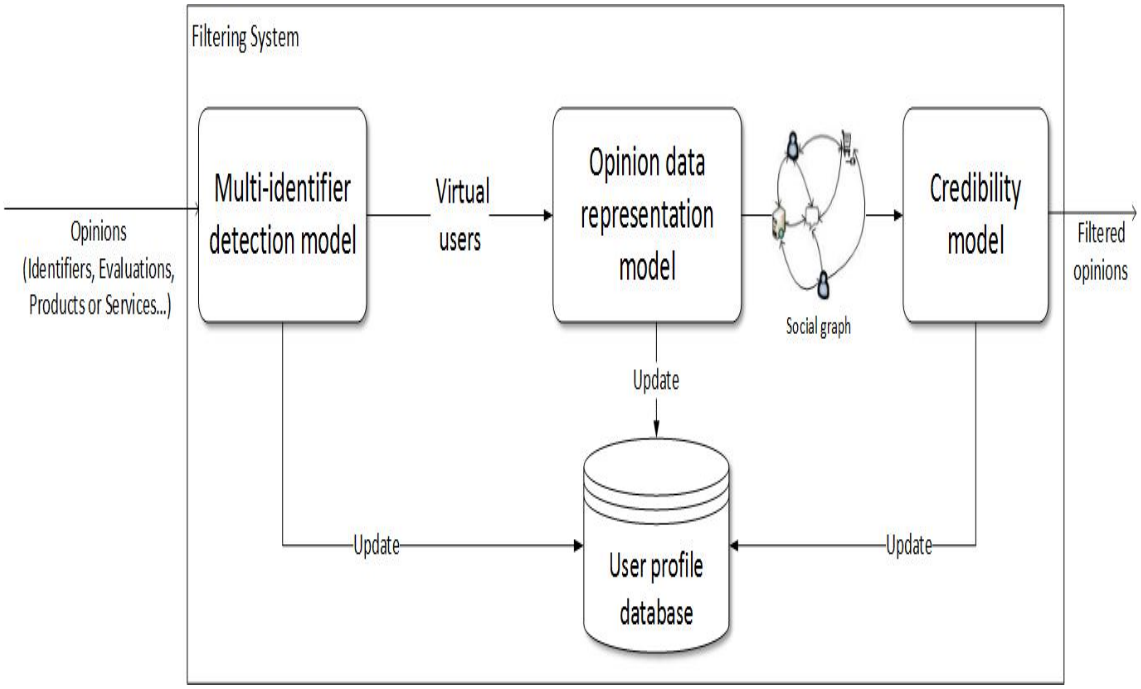
- We propose a model for detecting virtual users that correspond to the same physical user. The model combines different types of identifiers comparisons: comparison of email address, profile names, opinion publication dates and friends list. Each criterion taken alone produces a classification of the set of users. The technique of hierarchical agglomerative clustering (HAC) (William and Herbert, 1984 ; Daniel, 2013) is then used to reach a single virtual user classification. Each class corresponds to a set of virtual users that probably belong to the same physical user.

- We also propose a model for calculating users credibility. This credibility takes into account the users behavior and measures the consistency of opinions, the influence of the virtual environment of users, and the opinion of credibility that users have about one another.

- We propose an architecture for filtering opinions to improve their quality before calculating the product's reputation. This architecture is described in figure 3.2.1. The multi-identifiers detection model finds the identifiers that belong to the same physical user. Afterward, opinions data representation model ensures the elimination of opinions redundancy and produces as a result a social heterogeneous graph to structure users, products, evaluation sites and evaluations themselves. Finally, the credibility model affects a credibility score to reviewers by exploiting the principles of opinions consistency, inter-evaluation and inter-influence of users. After each step of the process, the user profile database is updated.

- We valid the proposed approach through experiments. This is based on a random data generation and a variation of different criteria considered in the detection of the true physical users. Our models result is a collection of opinions filtered to eliminate all forms of redundancy, and qualified with a weight representing the credibility of users. This credibility can be refined by working on the content of the opinions and comparing them to one another such as presented in the paper of Zohra *et al.* (2014).

Figure 3.2.1: General architecture of the opinions filtering system



3.3 Multi-identifiers detection model

3.3.1 Motivations

It is common that some users in social networks create multiple and different accounts. There are many reasons for doing this. For example, a user may post evaluations with different userids to advertise a product and make it more popular. Or to demote a competitor's product. Also he can simply belong to different social networks and websites (Facebook, Twitter, YouTube ...) without any malicious goals. Hence, detecting aliases belonging to the same physical user has become an urgent task.

In our approach, we take into account the existence of multiple independent Web sites that offer the possibility to evaluate products and services. The same product can be evaluated on different sites. We also consider the fact of the presence of users in different social networks.

3.3.2 Detection criteria

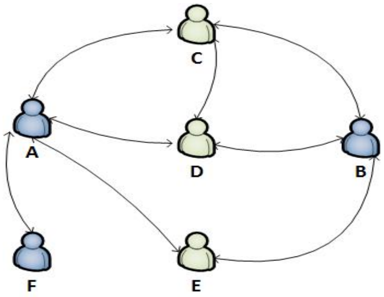
When trying to identify userids that belong to the same user, many features can be considered like profile names, but this becomes insignificant when working on malicious users who can provide very dissimilar user names for different accounts. IP addresses may be used too for recognizing same users. Yet, an IP Address may locate a single computer or it may locate a computer network. Also, almost all big organizations have their own private network that sits behind a firewall. They may use static or dynamic IP Addresses.

Taking into account these factors, the most obvious conclusion is that using only one criterion is not sufficiently significant to be qualified as a personally-identifiable data. We propose a model for detecting multiple userids based on a flexible combination of different criteria presented below :

1. **E-mail address.** Users generally create accounts in different social Web sites (Facebook, Twitter ...) using the same e-mail address. Especially in the case of non-malicious users. This criterion is very important but remains insufficient alone because it cannot detect multi-identifiers with different email addresses.
2. **Profile name.** In the non malicious case, users generally provide the same profile name for different social Web sites. Muniba *et al.* (2011) focus on finding similarities in usernames based on orthographic variations to detect aliases. This criterion has been taken into consideration in the work of Frederik*et al* (2013) also with the use of the Jaro-Winkler distance (William, 1990) to compare users names.
3. **Opinion publication dates.** Users tend to post evaluations from different accounts within a short time period. The Euclidean distance has been used to calculates how far away two time profiles are from each other in the approach of Frederik*et al.* (2013).
4. **Friends list.** Users mostly have the similar friends lists in different social

networks. Figure 3.3.1 shows social networks where A and B have three similar neighbors (C, D and E).

Figure 3.3.1: Example of a social network



3.3.3 Description of the multi-identifiers detection model

The detection of multiple identifiers consists in taking one or more criteria and generating classes of identifiers. Each class represents the identifiers that probably correspond to the same physical user. Criteria listed above, taken alone, produce each one different classifications. Two identifiers can belong to the same user according to one criterion and to different virtual user according to another. Algorithms for generating these classifications are not necessarily the same for all the criteria. We describe below the principles used for the generation of classifications and show how we combine them to arrive at a single final classification.

3.3.3.1 Detection of multi-identifiers with email address/profile name criterion

The algorithm for detection of multi-identifiers by email address criterion (respectively profile name) is simple and involves creating a class for all the identifiers sharing the same email (respectively profile name). The number of classes generated by each of these criteria is not the same, and the population of classes is not the same either.

3.3.3.2 Detection of multi-identifiers with opinion publication dates criterion

In this criterion, we consider two opinions having the same date of publication if the difference in dates does not exceed a certain time called threshold. Detection of multi-identifiers via this criterion is different from what previously defined because each identifier may publish many opinions while he has only one profile name and an email address.

Therefore, the principle of multi-identifiers detection via publication dates consists first, in comparing the dates between each pair of identifiers i and j then to calculate a similarity score ω_{ij} between identifiers i and j. The formula (1) indicates the calculating mode of the score ω_{ij}, where X represents the number of publications in a common time interval for two identifiers i and j ; variables P_i and P_j represent respectively the total number of publications for i and j.

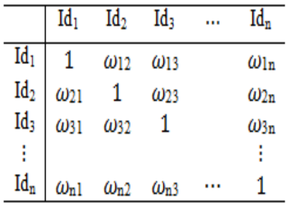
$$\omega_{ij} = \frac{2X}{|P_i + P_j|}\ , 0 \leq \omega_{ij} \leq 1 \tag{3.3.1}$$

It is important to note that the higher the number of posts in common dates are, the more probable that the two identifiers are coming from the same physical user.

The set of similarity scores between couples of identifiers is represented in a similarity matrix as described in figure 3.3.2. This matrix is then used via techniques of hierarchical agglomerative clustering (HAC) (William and Herbert, 1984 ; Daniel, 2013) to generate classes of similar identifiers as indicated in figure 3.3.3.

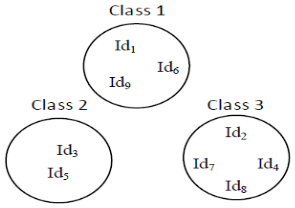
The hierarchical classification is justified by the fact that an identifier may be similar to many others who do not necessarily belong to the same class and with different degrees.

Figure 3.3.2: Similarity matrix



	Id_1	Id_2	Id_3	⋯	Id_n
Id_1	1	ω_{12}	ω_{13}		ω_{1n}
Id_2	ω_{21}	1	ω_{23}		ω_{2n}
Id_3	ω_{31}	ω_{32}	1		ω_{3n}
⋮					⋮
Id_n	ω_{n1}	ω_{n2}	ω_{n3}	⋯	1

Figure 3.3.3: Example of similar identifiers classes



3.3.3.3 Detection of multi-identifiers with friends list criterion

The criterion friends list is used in the same manner as the publication dates simply because each identifier can have many friends.

3.3.3.4 Multi-criteria detection of multi-identifiers

Detection of multi-identifiers by simultaneous application of different criteria, produces four different classifications of identifiers. It is about using these four classifications to produce ultimately only one. To do this, we use the HAC algorithm to combine different classes in one classification. This classification method is automatic and used in data analysis from a set of individuals. Its purpose is to classify individuals with similar behavior by a similarity criteria defined in advance. The most similar individuals will be put together in homogeneous groups.

The classification is agglomerative because it starts from a situation where all individuals are put alone in one class; it is hierarchical because it produces classes increasingly large. The method assumes that we have a similarity measure between individuals. It is then to create a similarity matrix but with the combination of the proposed features.

The similarity computation within the different criteria is calculated as the formula 3.3.2 shows, the Boolean function *Critn* verify whether identifiers i and j belong to the same user according the criteria n in which case it returns 1, otherwise it returns 0. Coefficients a, b, c et d correspond to the weights associated to the criteria.

$$Similarity_{ij} = a \times Crit1_{(i,j)} + b \times Crit2_{(i,j)} + c \times Crit3_{(i,j)} + d \times Crit4_{(i,j)} \quad (3.3.2)$$

$$a, b, c, d \in [0, 1] \, , a + b + c + d = 1$$

3.4 Opinion data representation model

The detection of multi-identifiers has an important impact on the representation of opinion data. We mean by opinion data all the data related to users, sites of evaluations, products and services, and evaluations themselves.

We propose to represent the opinion data first by an aggregation of identifiers that correspond to the same physical user and then as a social graph linking users, sites, products and evaluations.

3.4.1 Aggregation of similar identifiers and elimination of redundancy

In this step we aggregate all identifiers belonging to the same class in a single virtual user. This aggregation is to create a virtual identifier and assign to it all non-identical evaluations. Redundant evaluations are deleted (they correspond to having the same evaluation of the same product on different sites from the same user in a given period of time).

3.4.2 Representation of opinion data with the social graph

Initial users or/and aggregated users, their evaluations, web sites on which they asses products and products themselves are represented in one heterogeneous social graph.

Four nodes are defined : User, Evaluation, Product and Web Site. These nodes are connected by the relationships as shown in Figure 3.4.1.

Edges between nodes represent relationships connecting them. Examples of relations are:

A. Users evaluate products

B. Products are displayed on websites.

C. A user is connected with other users.

D. Evaluations are displayed on websites.

E. A user can also evaluate credibility of peers.

F. Products are evaluated .

G. Users provide evaluations

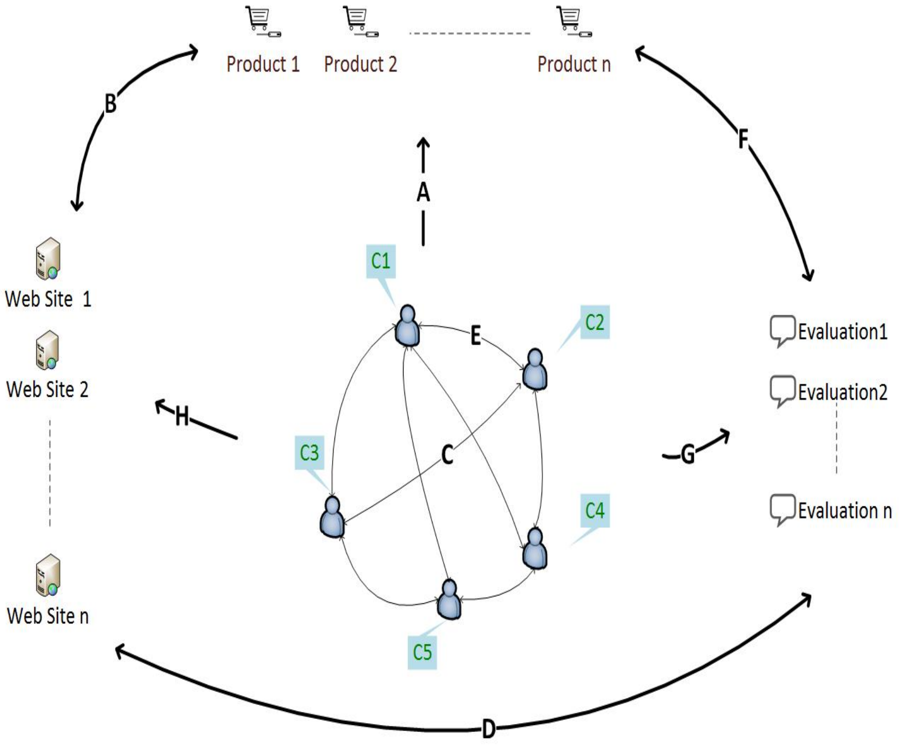
H. Users belong to different Web sites

The idea of a heterogeneous social graph was proposed in the paper of Guan *et al.*

(2011) but the authors do not represent the evaluators as an interconnected network.

Each user has a credibility score set to 0.5 and updated throughout the evaluation process. The following section will detail the calculation of the users credibility.

Figure 3.4.1: Social graph of opinions data



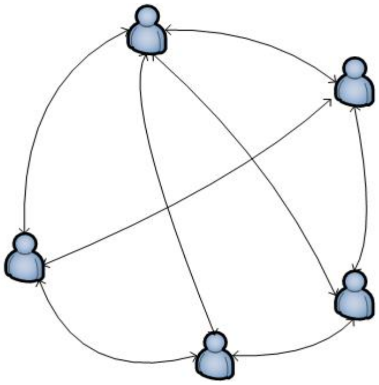
3.5 Credibility model

Interactions observed in the social graph between the different nodes allow to identify a number of principles for the calculation of the users credibility.

- **Inter-evaluation.** Users are able to evaluate and judge the credibility of peers. A user is credible if other users find him credible. Thus, The credibility score of a user U_i is presented by the formula 3.5.1 where $Numbre_{Ev}$ is the total number of evaluations about the user U_i, $Scale_sup$ is the upper limit of the scale considered (it is for example 5 for a scale of 1 to 5).

$$Score_{Ev}(U_i) = \frac{\sum Evaluations}{Number_{Ev} \times Scale_{sup}} \tag{3.5.1}$$

Figure 3.5.1: Inter-evaluation



- **Influence.** A user can be influenced by other users in his judgment. This influence can be direct or indirect through mutual friends. Figures 3.5.2 and 3.5.3 illustrate these two types of influence. We consider a user as credible if he is not influenced by other users.

Figure 3.5.2: Direct Influence

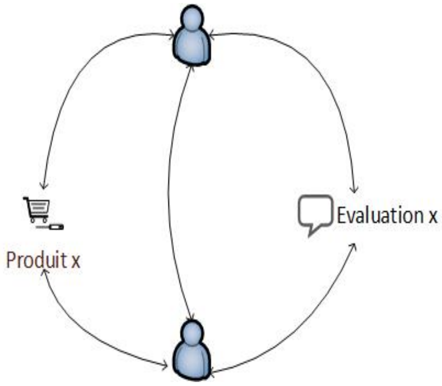
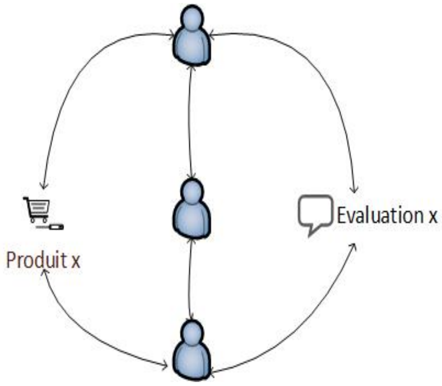


Figure 3.5.3: Indirect Influence



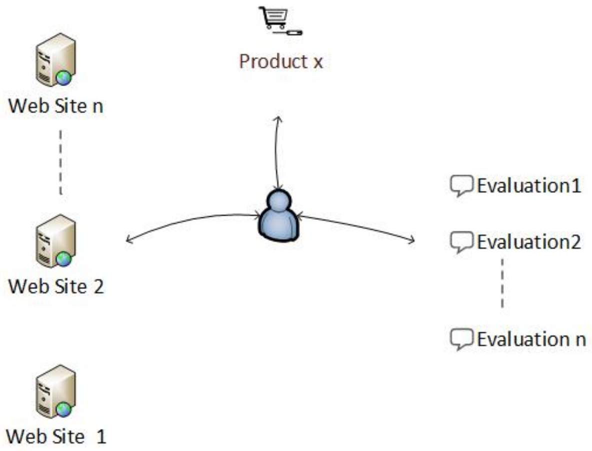
Therefore, the credibility score of a user U_i is expressed by the formula 3.5.2 where x and y denote the direct influence and indirect influence. These two variables take the value 1 when the relationships they represent exist, otherwise they are set to 0. Coefficients α and β represent the weights we would like to give to these two types of relationships.

$$Score_{inf}\left(U_i\right) = \frac{\alpha x + \beta y}{x + y}\,,\alpha \leq \beta\,, x, y \in \{0, 1\} \tag{3.5.2}$$

- **Consistency.** A user is consistent if he does not give different evaluations of the same product in one or more websites in a given period of time. The credibility score of a user U_i is presented by the formula 3.5.3 where Max_{Ev} is the maximal value of evaluations, Min_{Ev} is the minimal value and $Scale_s up$ is the upper limit of the scale considered.

$$Score_{Const}\left(U_i\right) = 1 - \frac{Max_{Ev} - Min_{Ev}}{Scale_{sup}} \tag{3.5.3}$$

Figure 3.5.4: Case of inconsistency



- **Computation of final user credibility.** The credibility of a user according to the above principles is given by the formula 3.5.4.

$$Credibility\,(U_i) = \frac{Score_{Ev} + Score_{Inf} + Score_{Const}}{3} \tag{3.5.4}$$

3.6 Conclusion

We detailed in this chapter the principles of our proposed approach: Opinion filtering. The system has 3 models: Multi-identifiers detection, opinion data representation and credibility model.

In order to evaluate the performance of our method, the next chapter will describe the implementation of our system and the the results generated by the framework besides the evaluation of its effectiveness.

Chapter 4

Experiments and results

4.1 Introduction

In this chapter, we present the implementation of our solution and the different modules that compose the framework.

We present the algorithms adopted and the tools used for the implementation. Finally, we present the results and the evaluation.

Data representation model and credibility model are not evaluated because, first, the data representation model is used to reduce redundancy which we can't evaluate and second, the credibility model is able to measure credibility based on some features but unfortunately there are no data available with the criteria proposed. The comparison of these two models performance and other models isn't possible in this case.

4.2 Algorithms

Based on our system architecture described in the previous chapter, we illustrate in this section the algorithms used for our system.

4.2.1 Detection of multi-identifiers with the criterion email-address

The algorithm 4.1 involves creating a class $Class_{email}$for all the identifiers sharing the same email address.

Algorithm 4.1 Email-Address

```
For each $email \in EMAIL$ do
    $Class_{email} \leftarrow \{\}$
    For each $id_i \in ID$ do
        If $email = id_i.email$ then
            $Class_{email} \leftarrow id_i + \{Class_{email}\}$
    end for
end for
```

4.2.2 Detection of multi-identifiers with the criterion profile name

Algorithm 4.2 works as the same way as the email address algorithm. The number of the classes generated by each of these criteria is not the same, and the population of classes is not the same either.

Algorithm 4.2 Profile Name

```
For each $pname \in PNAME$ do
    $Class_{Pname} \leftarrow \{\}$
    For each $id_i \in ID$ do
        If $pname = id_i.pname$ then
            $Class_{Pname} \leftarrow id_i + \{Class_{Pname}\}$
    end for
end for
```

4.2.3 Detection of multi-identifiers with the criterion opinion publication dates

As we described in the previous chapter, for this criterion, we need to produce the similarity matrix between each couple of identifiers. Algorithm 4.3 calculates this similarity score based on formula 3.3.1 from section 3 in chapter 3. After that, we use the algorithm of hierarchical agglomerative classification (HAC) described in algorithm 4.4 which gives us classes of the similar identifiers after creating the similarity matrix.

Algorithm 4.3 Publication dates similarity

```
$X \leftarrow 0$,
For each $t_i \in T_i$ , $1 \leq t_i \leq P_i$ do
    For each $t_j \in T_j$ , $1 \leq t_j \leq P_j$ do
        If $t_i = t_j$ then
            $X \leftarrow X + 1$// Number of posts in common time
        $\omega_{ij} = \frac{2X}{|P_i + P_j|}$
    end for
end for
```

Algorithm 4.4 HAC Algorithm

```
For $i = 1$ to $id.length$ do
    $class.add\left(newclass\left(id\left[i\right]\right)\right)$;
end for
While $class.length < nb.class$ do
    // Matrix creation
    $matSim = newMatrix\left(class.length, class.length\right)$;
    For $i = 1$ to $class.length$ do
        For $j = i + 1$ to $class.length$ do
            $matSim\left[i\right]\left[j\right] = Similarity\left(class\left[i\right], class\left[j\right]\right)$;
        end For
    end For
    // Max Similarity
    Let $(i, j)$as $matSim\left[i\right]\left[j\right] = max\left(matSim\left[k\right]\left[l\right]\right)$ with1 $\leq k \leq class.length$ and $k + 1 \leq l \leq class.length$ ;
    // Fusion of $class\left[i\right]$ and $class\left[j\right]$
    For all elements in $class\left[j\right]$ do
        $class\left[i\right].add\left(elements\right)$;
    end For
    $delete\left(class\left[j\right]\right)$;
end while.
```

4.2.4 Detection of multi-identifiers with the criterion friends list

This criterion is used as the same manner as the publication dates. This is due to the fact that both criteria present many variables (publication dates or friends) for each identifier. Algorithm 4.5 calculates the similarity between identifiers based on the criterion friends list. Then the HAC algorithm described in algorithm 4.4 is used to generate the matrix and gives classes of identifiers.

Algorithm 4.5 Friends list similarity

```
1. $X \leftarrow 0$,
2. For each $f_i \in F_i$ , $1 \leq t_i \leq TF_i$ do // $TF_i$ is the total number of friends of the identifiers $i$
3. For each $f_j \in F_j$ , $1 \leq t_j \leq TF_j$ do // $TF_j$ is the total number of friends of the identifiers $j$
4. If $f_i = f_j$ then
5. $X \leftarrow X + 1$// Number of friends in common between identifiers $i$ and $j$
6. $\omega_{ij} = \frac{2X}{|TF_i + TF_j|}$
7. end for
8. end for
```

4.2.5 Multi-criteria detection of multi-identifiers

As we described in section 3, chapter 3, we calculate the similarity score between identifiers based on formula 3.3.2. Algorithm 4.6 calculates this similarity score and once done, algorithm 4.4 is used.

Algorithm 4.6 Multi-criteria Similarity

```
For each $\{id_i, id_j\},\ i \neq j$ do
    $[Crit1 \leftarrow 0]$; $[Crit2 \leftarrow 0][Crit3 \leftarrow 0][Crit4 \leftarrow 0]$
    For each $Class_{email} \in ClassMAIL$ do
        If $\{id_i, id_j\} \subseteq Class_{email}$ then
            $[Class1 \leftarrow 1]$
    For each $Class_{Pname} \in ClassPNAME$ do
        If$\{id_i, id_j\} \subseteq Class_{Pname}$ then
            $[Class2 \leftarrow 1]$
    For each $Class_{Ptime} \in ClassPTIME$ do
        If $\{id_i, id_j\} \subseteq Class_{Ptime}$ then
            $[Class3 \leftarrow 1]$
    For each $Class_{Friends} \in ClassFRIENDS$ do
        If $\{id_i, id_j\} \subseteq Class_{friends}$ then
            $[Class4 \leftarrow 1]$
    $Similarity_{ij} = a \times Crit1 + b \times Crit2 + c \times Crit3 + d \times Crit4$
    $a + b + c + d = 1$ and $a > b > c > d$ and $a, b, c, d \in [0, 1]$
```

4.3 Experiments

4.3.1 Data description

It is difficult to find real databases to evaluate the approach and the proposed algorithms. We therefore conducted a random data generation. The database used contains 1000 identifiers each having a profile name, an e-mail address, at least 200 friends and at least 100 dates of publications.

4.3.2 Evaluation metrics

This section describes the experiment run in the study to evaluate the performance of our approach using the standards Accuracy, Recall and F-score.

The **Accuracy** (also known as specificity) measures how often the system is correct when it detect a conversation. It is calculated by dividing the number of correct outputs (true positive, TP) by the total number of the outputs. The total number of the outputs is the number of correct outputs plus the number of incorrect ones (false positive, FP).

$$Accuracy = \frac{|TP||}{|TP| + |FP|} \tag{4.3.1}$$

The **Recall** (also known as sensitivity) measures how often the system correctly finds the right classes to output. It is defined as proportion of true positives against potential correct outputs. The total number of potential correct outputs is the number of correct output (true positive, TP) plus the count of objects that should have been output but where not (true negative, TN).

$$Recall = \frac{|TP|}{|TP| + |TN|} \tag{4.3.2}$$

The **F-score** is the harmonic mean between Accuracy and Recall:

$$F - score = \frac{(2 \times Accuracy \times Recall)}{(Accuracy + Recall)} \tag{4.3.3}$$

4.4 Results and evaluation

The results given in Table 4.1 demonstrate the significance of considering multi-criteria approach. The overall performance scores of our framework are 84.23% recall, 88.64% accuracy and 86.37% F-Score.

Table 4.1: The approach results for 1000 identifiers

	Accuracy (%)	Recall (%)	F-score (%)
Email address	53.52	42.61	47.44
Profile name	38.12	27.82	32.16
Publication dates	9.41	5.24	6.73
Friends list	29.72	21.52	24.96
Combination of 4 criteria	**88.64**	**84.23**	**86.37**
Email address and publication dates	78.39	72.36	75.25
Profile name and friends list	76.61	64.42	69.98

The results of the experiments are shown in figures 4.4.1 and 4.4.2 where we considered various criteria independently, and also a combination of criteria.

In general, the results obtained show that there is a decrease in accuracy when the number of users increases. We can also notice that the individual use of each criterion gives low accuracy results compared to the results obtained in the case of a criteria combination.

These results show that the criterion Email address taken individually always works better than the rest of the other criteria with an accuracy of 53% for 100 users, the accuracy decreases as the number of users increases but remains stable at 36% for 1000 users. The use of the criterion publication dates alone is insignificant and accuracy does not exceed 10% regardless of the number of users (figure 4.4.1).

The four criteria combined together give accuracy above 80% for up to 900 users and touches 88% for 100 users. It decreases slowly and stabilizes at 78% for 1000 users. The combination of pairs of criteria e-mail address and publication dates on one hand and profile name and friends list on the other hand gives good results compared to results of criterion taken individually and accuracy stabilized between 65% and 70% for 1000 users. We can also see that the use of the criterion publication dates in combination with the e-mail address becomes meaningful as the use of e-mail address alone gave accuracy results between 36% and 53% and reached 78% in combination with publication dates.

The results presented indicate that it is important to use different criteria for multi-identifiers detection.

Figure 4.4.1: Results of criteria taken individually

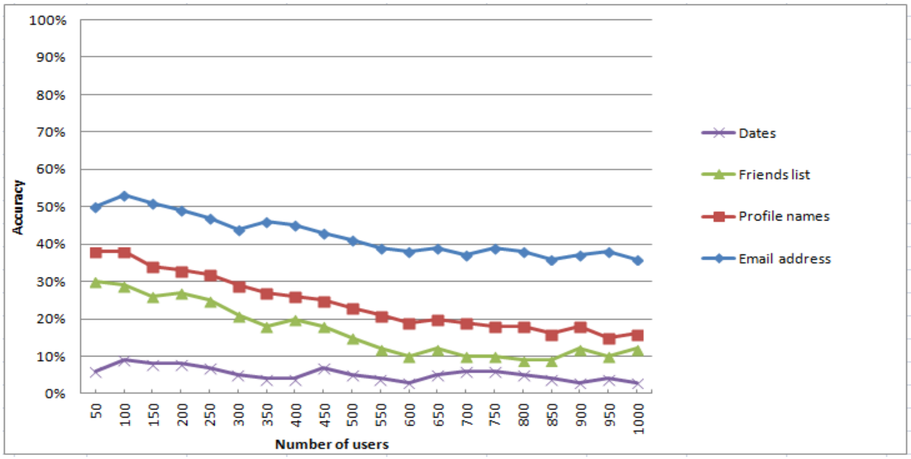
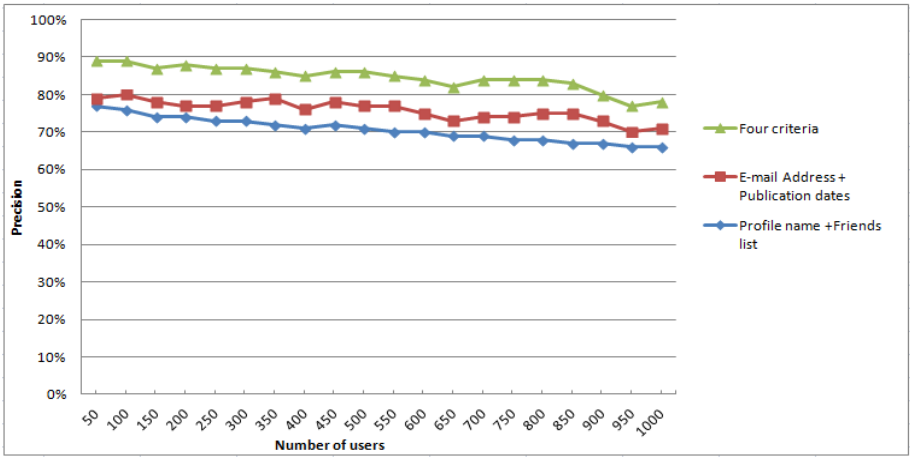


Figure 4.4.2: Results of different combinations of criteria



4.5 Conclusion

We exposed in this chapter in details our system implementation and its evaluation.

The models are constructed according to our framework described in the previous chapter.

Our experimental results have highlighted many interesting points. The fact that we combine a set of features to detect multi-identifiers improves the best accuracy. Furthermore, the comparison of our system with results of individual criterion using three metrics proved the higher performance of our search results confirming our model's effectiveness.

Conclusion

Online-resource reputation has become paramount. With the democratization of the Web and online social media (e.g., social networks and wikis) users have the opportunity to share their opinions on resources with millions of peers usually unknown. Unfortunately conflicting feedback and some users' malicious behaviors do not help develop concise opinions. This makes the task of measuring the reputation very difficult.

Therefore, it is important to filter opinions before any reputation measurement. In this work, we proposed an opinion filtering approach for social networks. This approach addresses limitations of existing systems like TIDY or Truth Finder such as lack of criteria combination and field restriction.

To achieve our framework's goals, we proceeded into three models:

The first is a model of virtual users detection corresponding to the same physical user. This model combines by using classification techniques, various criteria such as profile names and publication dates of opinions.

The second is opinion data representation model, it eliminates redundant opinions to produce a social heterogeneous graph that connects users, products, evaluation sites and evaluations together.

The third one is the credibility model which is characterized by calculating users credibility through a model based on the consistency of published opinions and influences that users exert on each other.

The experiments show that criteria combination leads to a quite interesting filtering. The four criteria combination accuracy reaches 88% for 100 users and stabilizes at 78% for 1000 users.

Interesting perspectives emerge to further strengthen the proposed approach. One of them is the use of theoretical models to represent and reason about uncertain information. Specifically, the use of probabilistic models by taking into account any form of uncertainty when detecting same users, calculating credibility, and producing probabilistic social graphs. The complexity of our problem will also be taken into consideration in future works. The approach should tend more to develop methods that give optimal solutions.

Bibliography

Anthony Etuk, Timothy J.Norman, Murat Sensoy, Chatschik Bisdikan, and Mudhakar Sribatsa. “TIDY: A Trust-Based Approach to Information Fusion through Diversity”. Information Fusion (FUSION), IEEE 16th International Conference. Pages 1188 - 1195. 9-12. Istanbul, July 2013.

Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos and Riddhiman Ghosh. “Spotting Opinion Spammers using Behavioral Footprints”. KDD’ 13 Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. Pages 632-640. NY, USA, 2013.

Danah m. boyd and Nicole B. Ellison. “Social Network Sites: Definition, History, and Scholarship”. Journal of Computer-Mediated Communication. Pages 210–230. October 2007.

Daniel Mullner. “Fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python”. Journal of statistical software. May 2013.

David Lazer, Alex S. Pentland, Lada Adamic, Sinan Aral, Albert L. Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler and Myron Gutmann. “Life in the network: the coming age of computational social science”. Journal: Science. New York, USA, 2009.

Dingding Wanga, Shenghuo Zhu b and Tao Li. “SumView: A Web-based engine for summarizing product reviews and customer opinions”. Expert Systems with Applications: An International Journal. Pages 27-33. Tarrytown, New York, USA, January, 2013.

Fayyad Usama, Piatetsky-Shapiro Gregory and Smyth Padhraic. “From Data Mining to Knowledge Discovery in Databases”. Book: Advances in knowledge discovery and data mining. Pages 1-34. Retrieved 17 December 2008.

Fredrik Johansson, Lisa Kaati and Amendra Shrestha. “Detecting Multiple Aliases in Social Media”. Proceedings of the 2013 IEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Pages 1004-1011. New York, USA, 2013.

Guan Wang, Sihong Xie, Bing Liu and Philip S. Yu. “Review Graph based Online Review Spammer Detection”. ICDM’11 Proceedings of the 2011 IEE International Conference on Data Mining. Pages 1242-1247. Washington, USA, 2011.

Hung-Ching Chen, Mark Goldberg, and Malik Magdon-Ismail. “Identifying Multi-ID Users in Open Forums”. Second Symposium on Intelligence and Security Informatics. USA, June 10-11, 2004.

Md Yusuf S Uddin, Md Tanvir Al Amin, Hieu Le, Tarek Abdelzaher, Boleslaw Szymanski andTommy Nguyen. “On diversifying Source Selection in Social Sensing”. 9th International Conference on Networked Sensing Systems, INSS’ 12. Pages 1-8. Belgium, June 11-14 2012.

Metzger Miriam J. “Making Sense of Credibility on the Web: Models for Evaluating Online Information and Recommendations for Future Research”. Journal of the American Society for Information Science and Technology. New York, USA, November 2007.

Mostafa Mohamed M. “More Than Words: Social Networks’ Text Mining for Consumer Brand Sentiments”. Expert Systems with Applications: An International Journal archive. Pages 4241-4251. Tarrytown, New York, USA, August 2013.

Muniba Shaikh, Nasrullah Memon and Uffe Koek Wiil. “Extended approximate string matching algorithms to detect name aliases”. Intelligence and Security Informatics (ISI), 2011 IEEE International Conference. Pages 216 – 219. Beijing, 10-12 July 2011.

Nicole B. Ellison, Charles Steinfield and Cliff Lampe. “The Benefits of Facebook “Friends:” Social Capital and College Students’ Use of Online Social Network Sites”. Journal of Computer-Mediated Communication. Pages 1143–1168, July 2007.

Nitin Jindal and Bing Liu. “Opinion Spam and analysis”. International Conference on Web Search and Web Data Mining, WSDM. New York, USA, 2008.

Oren Etzioni. “The world wide Web: Quagmire or gold mine”. Magazine: Communications of the ACM Issue 11. Pages 65-68. New York, USA, November 1996.

Raymond Kosala and Hendrik Blockeel. “Web mining Research: A Survey”. SIGKDD Explorations - Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining. New York, USA, June 2000.

Tieyun Qian and Bing Liu. “Identifying Multiple Userids of the same Author”. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing EMNLP. Pages 1124-1135.Washington, USA, October 18-21, 2013.

Tracy Riggs and Robert Wilensky. “An algorithm for automated rating of reviewers”. Proceedings of the First ACM/IEEE-CS Joint Conference on Digital libraries (JCDL). Pages 381-387. New York, USA 2001.

William E. Winkler. “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage”. Proceedings of the Section on Survey Research. Page 354-359. U.S. Bureau of the Census 1990.

William H. E. Day and Herbert Edelsbrunner. “Efficient Algorithms for Agglomerative Hierarchical Clustering Methods.” Journal of Classification. Pages 7–24. 1984.

Xin Luna Dong, Laure Berti-Equille and Divesh Strivastava. “Integrating conflicting data: the role of source dependence”. VLDB conference. Pages 550 – 661. August 2009.

Yi-Cheng Ku, Chih-Ping Wei and Han-Wei Hsiao. “To whom should I listen? Finding reputable reviewers in opinion-sharing communities”. Decision support systems. Pages 534–542. June 2012.

Zaki Malik and Athman Bouguettaya. “Reputation assessment for trust establishment among web services”. Very Large Data Bases (VLDB), 2009.

Zohra Saoud, Noura Faci, Zakaria Maamar and Djamal Benslimane. “A Fuzzy Clustering-based Credibility Model for Trust Management in a Service-oriented Architecture”. IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises. Parma, Italy, 2014.